Course One
Foundations of Data Science

Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. You can use this document as a guide to consider your responses and reflections at different stages of the data analytical process. Additionally, the PACE strategy documents can be used as a resource when working on future projects.

Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

- ☑ ~~Complete the PACE Strategy Document to plan your project while considering your audience members, teammates, key milestones, and overall project goal.~~
- ☑ ~~Create a project proposal for the data team.~~

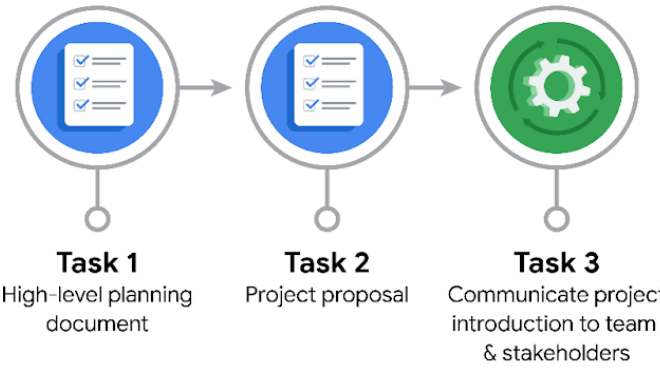
Relevant Interview Questions

Completing this end-of-course project will empower you to respond to the following interview topics:

- As a new member of a data analytics team, what steps could you take to get 'up to speed' with a current project? What steps would you take? Who would you like to meet with?
- How would you plan an analytics project?
- What steps would you take to translate a business question to an analytical solution?
- Why is actively managing data an important part of a data analytics team's responsibilities?
- What are some considerations you might need to be mindful of when reporting results?

Reference Guide

This project has three tasks; the following visual identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- Who is your audience for this project?

The audience for this project includes:

1. Internal Stakeholders:
   - Automatidata Data Analytics Team: Collaborators involved in model development, who need updates on tasks and timelines.
   - Deshawn Washington (Data Analysis Manager): Oversees the project and expects progress reports and milestone tracking.

2. External Clients:
   - New York City Taxi and Limousine Commission (TLC): The client focused on how the regression model will improve fare predictions and support decision-making.

- What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger needs of the client?

The goal is to build a regression model that accurately predicts taxi fares based on factors like distance, time of day, and other relevant variables. This model will help improve fare transparency, optimize pricing, and enhance customer trust.

The anticipated impact on the larger needs of the client, the New York City Taxi and Limousine Commission (TLC), includes better fare management, improved operational efficiency, and data-driven insights that can support policy-making and pricing regulations. This will ultimately help TLC make more informed decisions and enhance rider satisfaction.

- What questions need to be asked or answered?

The following key questions need to be asked or answered:

1. Data Questions:
   - What specific data has been collected by the TLC, and is it clean and ready for analysis?
   - Are there any missing values or outliers in the dataset that need to be addressed?
   - What additional variables might affect fare prices (e.g., traffic conditions, weather)?

2. Modeling Questions:
   - What type of regression model will be most effective for predicting fares?
   - How will we handle categorical variables like time of day or location?
   - What evaluation metrics should we use to assess model accuracy?

3. Client Questions:
   - What specific goals does the TLC have for this fare prediction model (e.g., accuracy, ease of use)?
   - Are there legal or regulatory constraints we need to consider?
   - How frequently will the model need to be updated as new data becomes available?

4. Operational Questions:
   - What are the key milestones and deadlines for the project?
   - Who will be responsible for maintaining and updating the model once deployed?
   - How will the model be integrated into the TLC's existing systems and operations?

- What resources are required to complete this project?

The following resources are required to complete this project:

1. Data Resources:
   - Access to the New York City Taxi and Limousine Commission (TLC) dataset, including historical fare, distance, and trip data.
   - Additional data sources that may include traffic patterns, weather data, and events that could impact taxi usage.

2. Technical Resources:
   - Data analysis and modeling tools (e.g., Python, R, or specialized software like Tableau or Power BI).
   - Computing resources, such as cloud services (e.g., AWS, Google Cloud) or local servers, for data storage and model training.

3. Human Resources:
   - Data analysts and data scientists to conduct analysis and develop the regression model.
   - Project manager to oversee timelines, milestones, and communication among team members.

4. Documentation and Communication Resources:
   - Tools for documentation (e.g., Google Docs, Confluence) to track project progress and findings.
   - Communication tools (e.g., Slack, email) for team collaboration and updates with the TLC.

5. Financial Resources:
   - Budget for potential software licenses, cloud services, and any additional data acquisition if necessary.

These resources will help ensure the successful completion of the project and support effective collaboration among team members.

- What are the deliverables that will need to be created over the course of this project?

The following deliverables will need to be created over the course of this project:

1. Data Cleaning and Preparation Report:
   Documentation outlining the data cleaning process, including handling of missing values and outliers.

2. Exploratory Data Analysis (EDA) Report:
   A report summarizing key insights from the dataset, visualizations, and initial findings that inform model development.

3. Regression Model:
   The final regression model, including code and documentation on how it was built, validated, and tested.

4. Model Evaluation Report:
   A report detailing the model's performance metrics, such as accuracy, RMSE (Root Mean Square Error), and other relevant evaluation criteria.

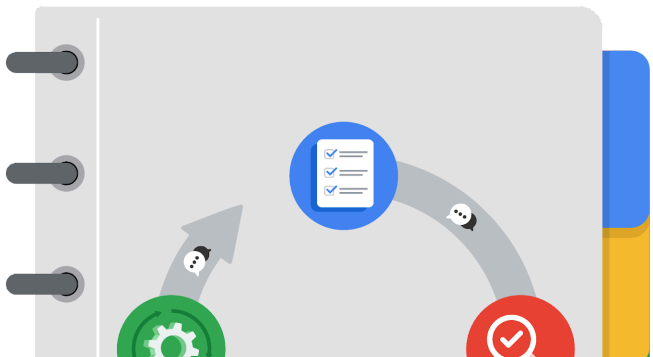
5. User Guide/Documentation:
   A comprehensive guide for stakeholders on how to use the regression model, interpret results, and update the model with new data.

6. Presentation to Stakeholders:
   A presentation summarizing the project findings, insights, and recommendations for the New York City Taxi and Limousine Commission (TLC).

7. Project Milestone Updates:
   Regular updates documenting progress against milestones, challenges faced, and adjustments made to the project plan.

These deliverables will provide a clear framework for project execution and ensure that stakeholders are informed throughout the process.

THE PACE WORKFLOW



[Alt-text: The PACE Workflow with the four stages in a circle: plan, analyze, construct, and execute.]

You have been asked to demonstrate for the company's data team how you would use the PACE

workflow to organize and classify tasks for the upcoming project. Select a PACE stage from the dropdown buttons. A few tasks involve more than one stage of the PACE workflow. Additionally, not every workplace scenario will require every task. Refer back to the Course 1 end-of-course portfolio project overview reading if you need more information about the tasks within the project.

Project tasks

Following are a group of tasks your company's data team has determined need to be completed within this project. The data analysis manager has asked you to organize these tasks in preparation for the project proposal document. First, identify which stage of the PACE workflow each task would best fit under using the drop down menu. Next, give an explanation of why you selected the stage for each task. Review the following readings to help guide your selections and explanation: The PACE stages and Communicate objectives with a project proposal. You will later reorder these tasks within a project proposal.

1. **Evaluating the model:** Construct

   Why did you select this stage for this task?

   I selected the "Construct" stage for this task because evaluating the model involves analyzing its performance and determining how well it meets the project objectives. During this stage, the team will assess metrics such as accuracy, precision, recall, and RMSE (Root Mean Square Error) to validate the model's effectiveness. This process helps in refining the model by identifying areas for improvement and ensuring it is ready for deployment or further testing.

2. **Conduct hypothesis testing:** Analyze **and** Construct

   Why did you select these stages for this task?

   I selected the "Analyze" stage for this task because hypothesis testing is a critical part of the analytical process where we evaluate assumptions about the data. In this stage, statistical tests are performed

to determine whether there is enough evidence to support or reject a hypothesis. This analysis is essential for understanding relationships between variables and validating the regression model being developed.

In addition to the "Analyze" stage, I also selected the "Construct" stage because the results of hypothesis testing inform the construction of the model itself. The findings will guide decisions about which variables to include or exclude in the final model, helping to build a more accurate and robust regression model based on the data analysis conducted.

3. **Begin exploring the data:** Plan

Why did you select this stage for this task?

I selected the "Plan" stage for this task because beginning the exploration of the data is a crucial initial step in the data analysis process. In this stage, the team outlines the approach for analyzing the dataset, including identifying key variables, understanding data distributions, and recognizing patterns or trends. This planning helps set the foundation for the analysis by ensuring that the team knows what to look for during the exploratory data analysis (EDA) phase and prepares for subsequent tasks in the project workflow.

4. **Data exploration and cleaning:** Analyze **and** Construct

Why did you select these stages for this task?

I selected the "Analyze"stage for this task because data exploration involves examining the dataset to uncover insights, patterns, and relationships between variables. This analysis is essential for understanding the structure and quality of the data, identifying potential issues such as outliers, missing values, or incorrect entries that need to be addressed.

I also selected the ” Construct” stage because cleaning the data is a preparatory step necessary for building the regression model. During this stage, the team will implement strategies to correct data quality issues identified during exploration, ensuring that the dataset is suitable for model development. This process is crucial to ensure the reliability and accuracy of the final model.

5. **Establish structure for project workflow (PACE):** Plan

Why did you select this stage for this task?

I selected the "Plan" stage for this task because establishing the structure for the project workflow involves outlining the steps and processes needed to successfully complete the project. This planning phase is essential for defining roles, responsibilities, timelines, and milestones. A clear workflow structure ensures that all team members understand their tasks and how they fit into the overall project, facilitating effective collaboration and project management throughout the project lifecycle.

6. **Communicate final insights with stakeholders:** Execute

Why did you select this stage for this task?

I selected the "Execute" stage for this task because communicating final insights involves presenting the results and findings to stakeholders in a clear and actionable manner. This stage is where the analysis and modeling efforts culminate in tangible outcomes that can be understood and utilized by decision-makers. Effective communication ensures that stakeholders are informed of the insights gained from the project and can make informed decisions based on the results. This step is crucial for translating the analytical work into practical applications that meet the needs of the client.

7. **Compute descriptive statistics:** Analyze

Why did you select this stage for this task?

I selected the "Analyze" stage for this task because computing descriptive statistics involves summarizing and interpreting the main features of the dataset. This analysis provides insights into the data's central tendencies, variability, and distribution patterns. By calculating measures such as mean, median, mode, standard deviation, and range, the team can better understand the data's characteristics, identify trends, and inform subsequent modeling decisions. This step is essential for laying the groundwork for deeper analysis and ensuring that the data is well-understood before proceeding to more complex analytical tasks.

8. **Visualization building: Analyze and Construct**

Why did you select these stages for this task?

I selected the "Analyze" stage for this task because building visualizations is an important part of the analysis process. Visualizations help to explore the data, identify patterns, and communicate insights effectively. This stage allows the team to analyze relationships between variables, trends over time, and other significant features of the dataset, enhancing understanding of the data.

I also selected the "Construct" stage because creating visualizations is crucial for constructing a compelling narrative around the data findings. The visualizations will be used to support the final model and present results to stakeholders, making it essential to build them accurately and effectively as part of the project deliverables. These visualizations are integral to the overall construction of the analysis and reporting process.

9. **Write a project proposal: Plan**

I selected the "Plan" stage for this task because writing a project proposal involves outlining the project's objectives, scope, tasks, and milestones. This stage focuses on creating a structured approach for the project, ensuring that all stakeholders understand the goals and expectations before the project begins. The proposal acts as a roadmap for the team, guiding the project's execution and aligning efforts with the client's needs. Thank you for pointing that out!

10. **Build a regression model: Construct and Execute**

Why did you select this stage for this task?

I selected the "Construct" stage for this task because building a regression model involves creating the analytical framework that will be used to analyze the data. This stage focuses on developing the model, selecting appropriate features, and applying statistical techniques to make predictions based on the data. It is a critical phase where theoretical concepts are translated into a tangible model that can be tested and validated.

I also selected the "Execute" stage because once the model is built, it must be implemented and run against the data to generate predictions. This execution phase involves evaluating the model's performance, interpreting results, and ensuring that it meets the project's objectives. Effective execution is essential for translating the model's construction into actionable insights that can inform decision-making.

11. **Compile summary information about the data:** Analyze

Why did you select this stage for this task?

I selected the "Analyze" stage for this task because compiling summary information involves examining the dataset to understand its structure, characteristics, and key features. This stage is crucial for generating insights from the data and identifying patterns or trends that may influence subsequent analysis. By summarizing information such as the number of records, variable types, and basic statistics, the team can assess data quality and readiness for further analysis or modeling. This foundational step is essential for guiding the project's direction and ensuring that the data is well-understood before proceeding to more complex tasks.

12. **Build machine learning model:** Construct

Why did you select this stage for this task?

I selected the "Construct" stage for this task because building a machine learning model involves creating the computational framework that will learn from the data. This stage focuses on selecting the appropriate algorithms, defining the model architecture, and training the model using the available data. It is a critical phase where theoretical concepts are implemented into a functioning model. Proper construction is essential to ensure that the model is capable of making accurate predictions and generalizing well to unseen data.